

ATYPICAL SCALING BEHAVIOR PERSISTS IN REAL WORLD INTERACTION NETWORKS

HARRY CRANE AND WALTER DEMPSEY

ABSTRACT. Scale-free power law structure describes complex networks derived from a wide range of real world processes. The extensive literature focuses almost exclusively on networks with power law exponent strictly larger than 2, which can be explained by constant vertex growth and preferential attachment. The complementary scale-free behavior in the range between 1 and 2 has been mostly neglected as atypical because there is no known generating mechanism to explain how networks with this property form. However, empirical observations reveal that scaling in this range is an inherent feature of real world networks obtained from repeated interactions within a population, as in social, communication, and collaboration networks. A generative model explains the observed phenomenon through the realistic dynamics of constant edge growth and a positive feedback mechanism. Our investigation, therefore, yields a novel empirical observation grounded in a strong theoretical basis for its occurrence.

Self-organizing dynamics of many processes produce a common heterogeneous structure characterized by power law degree distributions, which have been discovered in the World Wide Web [1–3], social networks [4], telecommunications networks [5], biological networks [6], and many others [7–10]. A network exhibits *power law* degree distribution with exponent $\gamma > 1$ if the proportion p_k of vertices with degree k satisfies $p_k \sim k^{-\gamma}$ for large k . Figure 1 plots the degree distributions of four well known networks: the actors collaboration network [1], Enron email network [11], Wikipedia voting network [12], and Facebook social circles network [13, 14]. The power law exponent in each of these networks is between 1 and 2, behavior that cannot be explained by preferential attachment models. Preferential attachment dynamics provide an intuitive description of networks that undergo constant vertex growth and exhibit power law greater than 2. These properties do not accurately describe the networks in Figure 1:

- (A) In the actor collaboration network, vertices correspond to actors and edges represent that two actors were cast in the same movie. Here the data permits multiple edges between vertices if the actors were cast together more than once. Thus, the network grows as a consequence

Date: July 13, 2015.

H. Crane is partially supported by NSF grant DMS-1308899 and NSA grant H98230-13-1-0299.

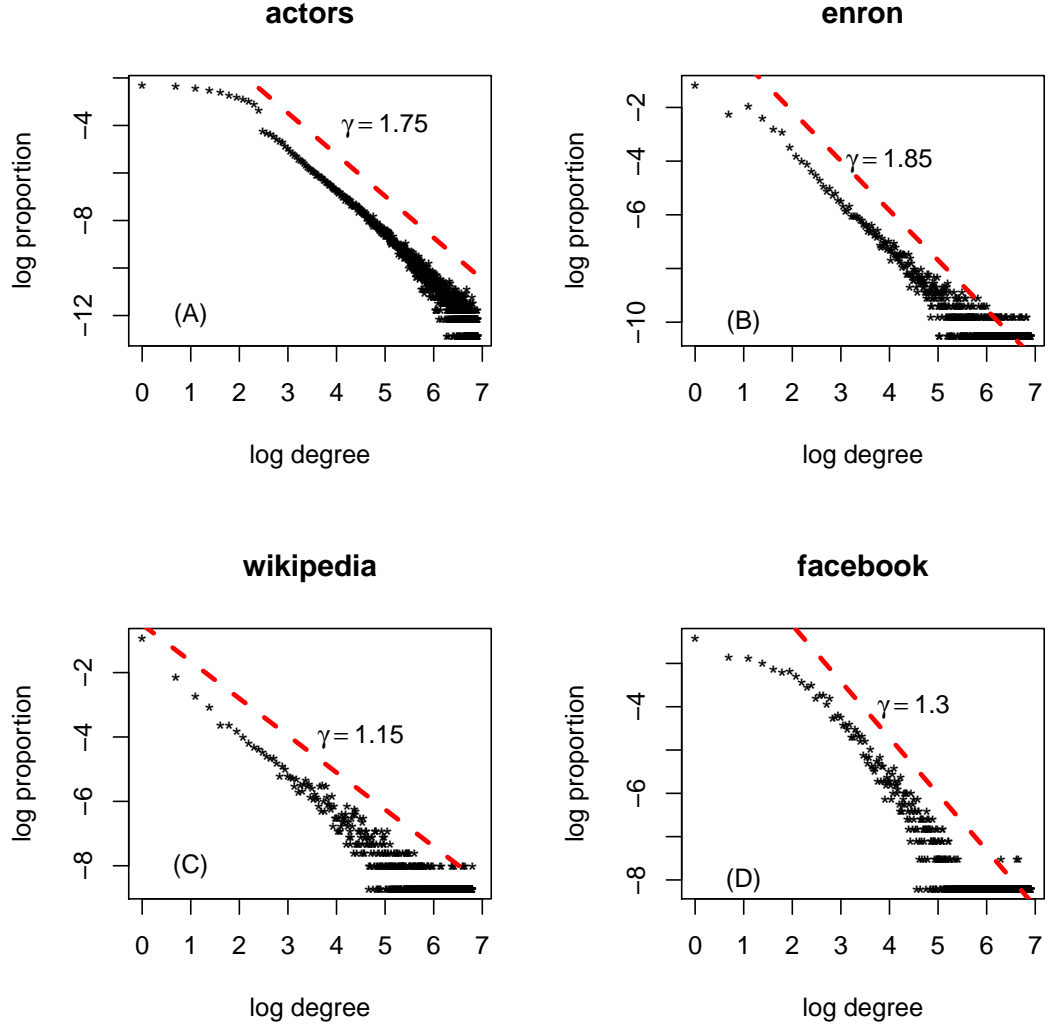


FIGURE 1. Empirical degree distributions for (A) Actors collaboration network, (B) Enron e-mail network, (C) Wikipedia link network, and (D) Facebook social circles network. In each panel, the slope of the dashed line is $-\gamma$, where γ is the estimated power law exponent. Fitting the two-parameter generative model in (1) to the data, we obtain estimates (A) $(\alpha_{\text{actor}}, \theta_{\text{actor}}) = (0.75, 1.14)$, (B) $(\alpha_{\text{enron}}, \theta_{\text{enron}}) = (0.85, -0.63)$, (C) $(\alpha_{\text{wiki}}, \theta_{\text{wiki}}) = (0.15, 350)$, and (D) $(\alpha_{\text{fb}}, \theta_{\text{fb}}) = (0.30, 168)$.

- of movie production, i.e., edge formation, rather than the influx of new actors.
- (B) In the Enron email network, vertices correspond to employees at the Enron corporation and an edge represents that an email has been exchanged between those employees. As emails are exchanged, new edges form without any requirement that new vertices be added.
 - (C) The Wikipedia voting network represents voting behavior for elections to the administrator role in Wikipedia. Vertices are Wikipedia users and a directed edge points from i to j if user i voted for user j . The network grows when elections are held, i.e., new edges are formed.
 - (D) The Facebook social circles network represents friendships among Facebook users. The network grows by the formation of new friendships, i.e., edges, which usually result from social interactions among users.

Each of the above networks grows by the addition of edges that connect according to a positive feedback mechanism, whereby past interactions reinforce future behavior. For example, an email sent from employee A to employee B is likely to be reciprocated by a reply from B to A; actors cast together in one movie likely play complementary roles which may be suitable in future movies; and so on.

While positive feedback exhibits obvious similarities to preferential attachment, it differs in that edge formation need not be accompanied by the addition of new vertices. Furthermore, the range of the power law exponent implies additional growth properties about the network. Power law exponent $\gamma > 2$ implies that the expected vertex degree grows at rate $\sum_{k=1}^n k \cdot k^{-\gamma} \approx \int_1^n x^{-\gamma+1} dx \sim O(1)$ as a function of the number of vertices, making the total number of edges grow at the rate $n \cdot O(1) = O(n)$. Therefore, preferential attachment models implicitly assume a network for which the number of edges grows linearly with the number of vertices. On the other hand, for $1 < \gamma < 2$, the expected degree grows at rate $\sum_{k=1}^n k \cdot k^{-\gamma} \sim O(n^{2-\gamma})$ as a function of the number of vertices n , indicating total edge growth at rate $n \cdot O(n^{2-\gamma}) = O(n^{3-\gamma})$ in the intermediate range between sparsity $O(n)$ and density $O(n^2)$.

Some recent progress in the mathematical literature demonstrates the fundamentally different structural properties of sparse and dense networks [15–17]. Figure 1 suggests that power law exponent between 1 and 2 is also of scientific interest, and understanding this intermediate range should provide important insights into the structure of real world networks. We replicate these features in the following generative model, which produces scale-free networks with exponent $1 < \gamma < 2$ and closely resembles how networks (A)–(D) form. We generate a network with n edges by sequentially adding one edge at each time $t = 1, 2, \dots, n$. Our model is determined by two parameters α and θ in the range $0 < \alpha < 1$ and $\theta > -\alpha$. Before time t ,

the network has $t - 1$ edges and a random number of vertices N_t , with the initial condition $N_1 = 0$. We label these vertices $i = 1, \dots, N_t$ and write $D(i, t)$ to denote the total degree of vertex i before the t th edge is added. (Note that each self-loop from a vertex to itself contributes 2 to its degree.) When the t th edge arrives, its two incident vertices $v_1(t), v_2(t)$ are chosen randomly among vertices $1, \dots, N_t$ and a new vertex $N_t + 1$ as follows. With $N_t^1 = N_t$, we first choose $v_1(t)$ randomly with probability

$$(1) \quad \text{pr}(v_1(t) = i) \propto \begin{cases} D(i, t) - \alpha, & i = 1, \dots, N_t^1 \\ \theta + \alpha N_t^1, & i = N_t^1 + 1. \end{cases}$$

After choosing $v_1(t)$, we define N_t^2 according to whether or not $v_1(t)$ is a newly observed vertex: if $v_1(t) = N_t^1 + 1$, then we define $N_t^2 = N_t^1 + 1$; otherwise, we put $N_t^2 = N_t^1$. We then choose $v_2(t)$ as in (1) with N_t^1 replaced by N_t^2 . When generating a network with directed edges, we orient edges to point from $v_1(t)$ to $v_2(t)$; in the undirected case, the edge between $v_1(t)$ and $v_2(t)$ has no orientation. We write G_n to denote the network generated after n steps of this procedure.

The above generative model produces a sequence of networks $(G_n)_{n=1,2,\dots}$, where G_n has n edges and a random number of vertices N_n . For $k = 1, 2, \dots$, we write $N_n(k)$ to denote the number of vertices in G_n with degree k , so that $N_n = \sum_{k \geq 1} N_n(k)$. From properties of the generating mechanism in (1) [18], the empirical degree distributions $p_n(k) = N_n(k)/N_n$ converge to $\alpha \cdot k^{-(\alpha+1)}/\Gamma(1-\alpha)$, where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ is the gamma function. The simulation results in Figure 2 verify this property of our model. Moreover, the expected number of vertices satisfies

$$(2) \quad E(N_n) \sim \frac{\Gamma(\theta + 1)}{\alpha \cdot \Gamma(\theta + \alpha)} (2n)^\alpha, \quad \text{as } n \rightarrow \infty.$$

Given an observed power law exponent $1 < \gamma < 2$, we can use these two properties to estimate the model parameters α and θ by setting $\alpha = \gamma - 1$ and choosing the value of θ so that Equation (2) is satisfied by the observed network. The estimated parameters in the caption of Figure 1 were obtained by this method.

Remarkably, we can express the probability distribution of the random network G_n in closed form by:

$$(3) \quad \text{pr}(G_n = G) = \alpha^{\#V(G)} \frac{(\theta/\alpha)^{\uparrow \#V(G)}}{\theta^{\uparrow (2n)}} \prod_{v: \deg(v) > 1} (1 - \alpha)^{\uparrow (\deg(v)-1)}$$

where G is any network with n edges that can be generated by (1), $\deg(v)$ is the degree of vertex v in G , $\#V(G)$ is the number of vertices in G , and $x^{\uparrow j} = x(x+1)\cdots(x+j-1)$ is the ascending factorial function. A further important property of G_n is that its distribution (3) is independent of the order in which edges arrive during network formation. As this information is typically unavailable for network data, viable statistical models should be

agnostic to it. Nevertheless, many network models, including preferential attachment models, do depend on the order of arrival, severely limiting the scope of statistical inferences [19]. Under our model, this lack of information has no adverse consequences. Therefore, we expect that the discovery of (3) and its intuitive explanation of network formation should lead to significant progress in statistical network analysis.

Our generating mechanism allows for self-loops and multiple edges between vertices, features common in many of the interaction networks we consider. For the Enron and actors networks, respectively, self-loops correspond to emailing oneself and acting in the same movies as oneself, while multiple edges reflect an exchange of multiple emails between individuals and a casting of the same actors in multiple movies. Although these features may be present in the underlying real world phenomenon, network datasets are often simplified by reducing multiple edges to a single edge. In fact, of the four networks in Figure 1, only the actor collaboration network dataset records multiple edges. Thus, Figure 1 suggests that atypical scaling is not only present in interaction networks with multiple edges but also in their projection to a simple network by reducing multiple edges to a single edge. Figure 2 demonstrates that our model preserves the same scaling under this operation.

The parameters of our model have a clear interpretation in terms of the network generating mechanism. In (1), we see that α controls the rate at which a vertex accumulates edges, leading to the explicit relationship between α and the power law exponent $\gamma = \alpha + 1$. Given the value of α , θ controls the growth of vertices, with large values corresponding to faster growth. The θ parameter exhibits its biggest influence at the beginning of network formation. High estimates of θ for the Wikipedia and Facebook networks support the conclusion that most votes in Wikipedia elections involve users who did not participate in previous elections and the formation of Facebook social circles begins with rapid addition of new individuals. The moderate estimate of θ for the actors network supports the opposite conclusion; indeed, a core of the same movie actors are cast repeatedly while the majority of actors struggle for roles. The negative estimate of θ for the Enron network reflects the tendency for communication within a closed group to outpace the rate at which new team members are introduced.

The occurrence of power law exponent between 1 and 2 in several common network datasets brings forth a previously undetected feature of real world network evolution. While our sequential construction in (1) and preferential attachment dynamics both grow the network in a size-biased manner—higher degree vertices accumulate edges at a faster rate—network growth under our model is driven by the addition of edges, which accurately reflects the dynamics of the underlying network. Preferential attachment models, on the other hand, achieve the complementary power law behavior by sequential addition of vertices, behavior not reflective of networks (A)-(D). Even with state of the art methods [20, 21], statistical network models are

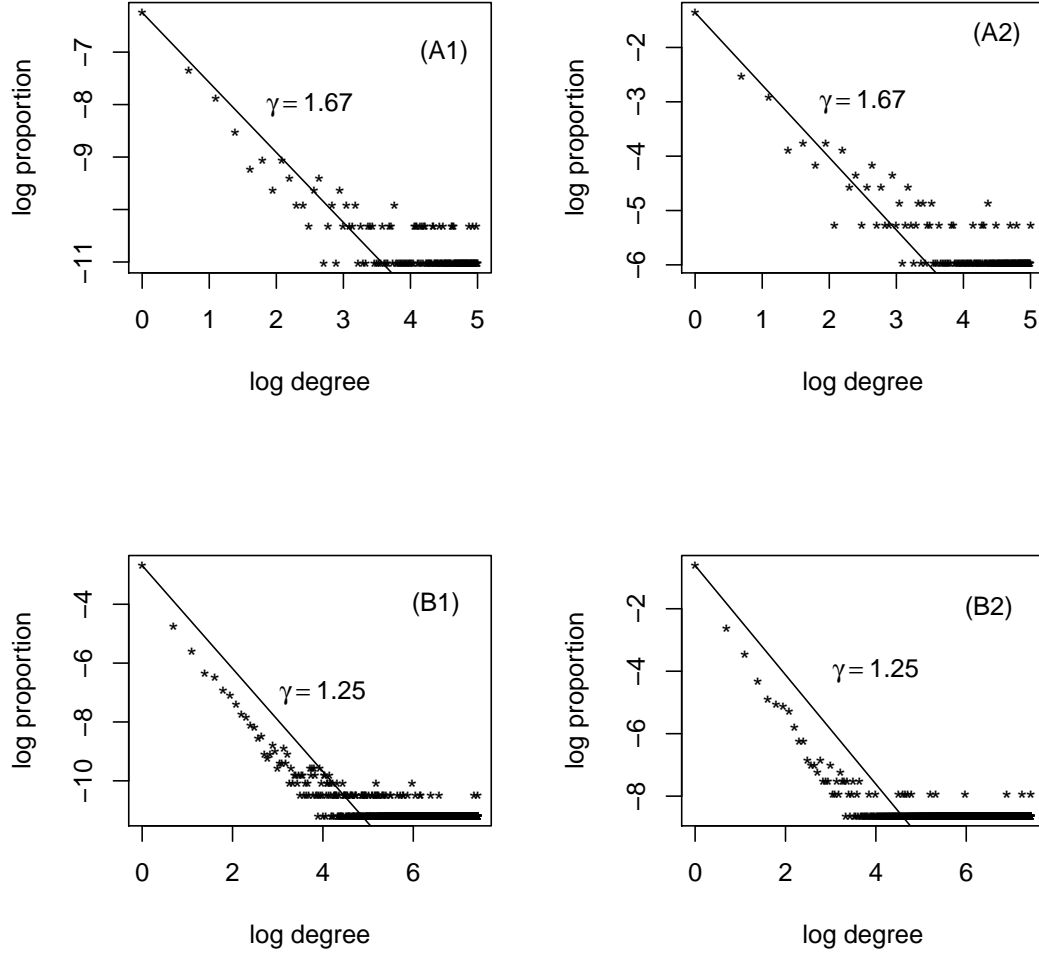


FIGURE 2. Simulation results showing degree distribution of networks and their projection to a simple network by removing multiple edges. (A1) Network generated from model with parameters $(\alpha, \theta) = (0.67, 1)$, (B1) Network generated from model with parameters $(\alpha, \theta) = (0.25, 1)$, (A2) Simple network obtained by reducing multiple edges to single edge in (A1) network, (B2) Simple network obtained by reducing multiple edges to single edge in (B1) network. Results suggest that the generated network and its induced simple network both exhibit power law of similar degree.

not sufficiently robust to answer many questions of practical interest [19]. Our model also possesses fundamental statistical properties that lead to straightforward estimation of the parameters α and θ and, hence, the power law exponent. Explicit calculation of the distribution (3) opens the door to much more detailed statistical analyses by likelihood-based and Bayesian techniques. Although power law exponent in this range has not received much attention, we expect that it is widespread in real world interaction networks. Our framework should lay the foundation for future investigations, both scientific and mathematical.

REFERENCES

- [1] Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). URL <http://dx.doi.org/10.1126/science.286.5439.509>.
- [2] Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. *ACM Comp. Comm. Review* **29** (1999).
- [3] Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. Trawling the web for emerging cyber communities. *Proceedings of the 8th World Wide Web Conference* (1999).
- [4] Watts, D. & Strogatz, S. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
- [5] Abello, J., Buchsbaum, A. & Westbrook, J. A functional approach to external graph algorithms. *Proceedings of the 6th European Symposium on Algorithms* 332–343 (1998).
- [6] Jeong, H., Mason, S., Barabási, A.-L. & Oltvai, Z. Lethality and centrality in protein networks. *Nature* **411**, 41 (2001).
- [7] Chung, F. & Lu, L. *Complex graphs and networks*, vol. 107 of *CBMS Regional Conference Series in Mathematics* (Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2006).
- [8] Dorogovtsev, S. N. & Mendes, J. F. F. *Evolution of networks* (Oxford University Press, Oxford, 2003). URL <http://dx.doi.org/10.1093/acprof:oso/9780198515906.001.0001>. From biological nets to the Internet and WWW.
- [9] Durrett, R. Some features of the spread of epidemics and information on a random graph. *Proceedings of the National Academy of Sciences* **107**, 4491–4498.
- [10] Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (electronic) (2003). URL <http://dx.doi.org/10.1137/S003614450342480>.
- [11] McAuley, J. & Leskovec, J. Learning to discover social circles in ego networks. *NIPS* (2012).
- [12] Leskovec, J., Huttenlocher, D. & Kleinberg, J. Signed networks in social media. *CHI* (2010).
- [13] Klimt, B. & Yang, Y. Introducing the enron corpus. *CEAS* (2004).
- [14] Leskovec, J., Lang, K., Dasgupta, A. & Mahoney, M. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* **5**, 29–123 (2009).
- [15] Borgs, C., Chayes, J. T., Cohn, H. & Zhao, Y. An L^p theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions. *accessed at arXiv:1401.2906* (2014).
- [16] Borgs, C., Chayes, J. T., Cohn, H. & Zhao, Y. An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. *accessed at arXiv:1408.0744* (2014).
- [17] Lovász, L. & Szegedy, B. Limits of dense graph sequences. *J. Comb. Th. B* **96**, 933–957 (2006).
- [18] Pitman, J. *Combinatorial stochastic processes*, vol. 1875 of *Lecture Notes in Mathematics*.

- [19] McCullagh, P. What is a statistical model? *Ann. Statist.* **30**, 1225–1310 (2002). URL <http://dx.doi.org/10.1214/aos/1035844977>. With comments and a rejoinder by the author.
- [20] Bickel, P. & Chen, A. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 21068–21073 (2009).
- [21] Bickel, P., Chen, A. & Levina, E. The method of moments and degree distributions for network models. *Ann. Statist.* (2011).

DEPARTMENT OF STATISTICS & BIostatISTICS, RUTGERS UNIVERSITY, 110 FRELINGHUYSEN AVENUE, PISCATAWAY, NJ 08854, USA
E-mail address: hcrane@stat.rutgers.edu
URL: <http://stat.rutgers.edu/home/hcrane>

DEPARTMENT OF STATISTICS, UNIVERSITY OF MICHIGAN, 1085 S. UNIVERSITY AVE, ANN ARBOR, MI 48109, USA
E-mail address: wdem@umich.edu